Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation

Barbara Befani and John Mayne

Abstract This article proposes a combination of a popular evaluation approach, contribution analysis (CA), with an emerging method for causal inference, process tracing (PT). Both are grounded in generative causality and take a probabilistic approach to the interpretation of evidence. The combined approach is tested on the evaluation of the contribution of a teaching programme to the improvement of school performance of girls, and is shown to be preferable to either CA or PT alone. The proposed procedure shows that established Bayesian principles and PT tests, based on both science and common sense, can be applied to assess the strength of qualitative and quali-quantitative observations and evidence, collected within an overarching CA framework; thus shifting the focus of impact evaluation from 'assessing impact' to 'assessing confidence' (about impact).

1 Introduction

It is becoming increasingly seen that for exploring causality in many real-world intervention settings, alternatives to the traditional counterfactual approaches need to be used, such as discussed in Stern *et al.* (2012) and Mayne and Stern (2013). Among the reasons for this are that many evaluations are conducted after the intervention is in place; ethical considerations limit the use of random assignment; interventions may be aimed at the entire population; or, with the increasing complexity of many interventions, setting up counterfactuals is not possible or practical.

Contribution analysis (Mayne 2012b, 2008, 2001) and process tracing (Beach and Pedersen 2011; Collier 2011; Bennett 2010, 2008) both seek to make causal inferences about cause and effect using non-counterfactual approaches based on similar analysis tools: causal mechanisms and theories of change. Contribution analysis (CA) comes out of the field of evaluation, while process tracing (PT) has emerged from the analysis of historical events. This article explores the relationship between the two methodology approaches and, using a hypothetical evaluation of a development intervention addressing girls' education, suggests a way to combine them. In particular, it explores using PT from an evaluation perspective, looks at how PT could be used to strengthen a CA, and identifies a combined CA-PT procedure for determining the contribution made to outcomes.

The identified procedure is not aimed at 'measuring impact' as such but rather at 'increasing our confidence' that the intervention had an impact. It illustrates the practical steps and consequences of grounding our thinking on Bayesian probability, which – compared to frequentist probability and statistics – offers more flexibility in terms of combining information from a variety of sources.

We first provide short overviews of perspectives on causality, contribution analysis and process tracing, before discussing the example and its evaluation from a combined perspective. We conclude that a CA framework is a useful entry point for the application of PT to impact evaluation, and that the application of PT principles and tests strengthens the conclusions

IDS Bulletin Volume 45 Number 6 November 2014
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

Table 1 **Four frameworks for causal inference**

	Aspect of causal relation	Causal question	Mill's methods	Description of causal mechanism
Counterfactual	Association between single cause and single effect	Did the intervention cause the effect? How much is the net effect of the intervention?	Difference	None
Regularity			Agreement, concomitant variation	None
Configurational	Association between configurations of conditions and effects; description of causal mechanism	What configurations of factors are necessary and/or sufficient for the effect?	Agreement and difference but only applied to causal packages	Only the basic ingredients are described: conditions, their combinations and disjunctions
Generative	Description of causal mechanism	How was the effect produced? How did it come about?	None	In-depth

Source Adapted from Stern *et al.* (2012).

reached in CA by linking the CA process more directly with an established method.

1.1 Perspectives on causality

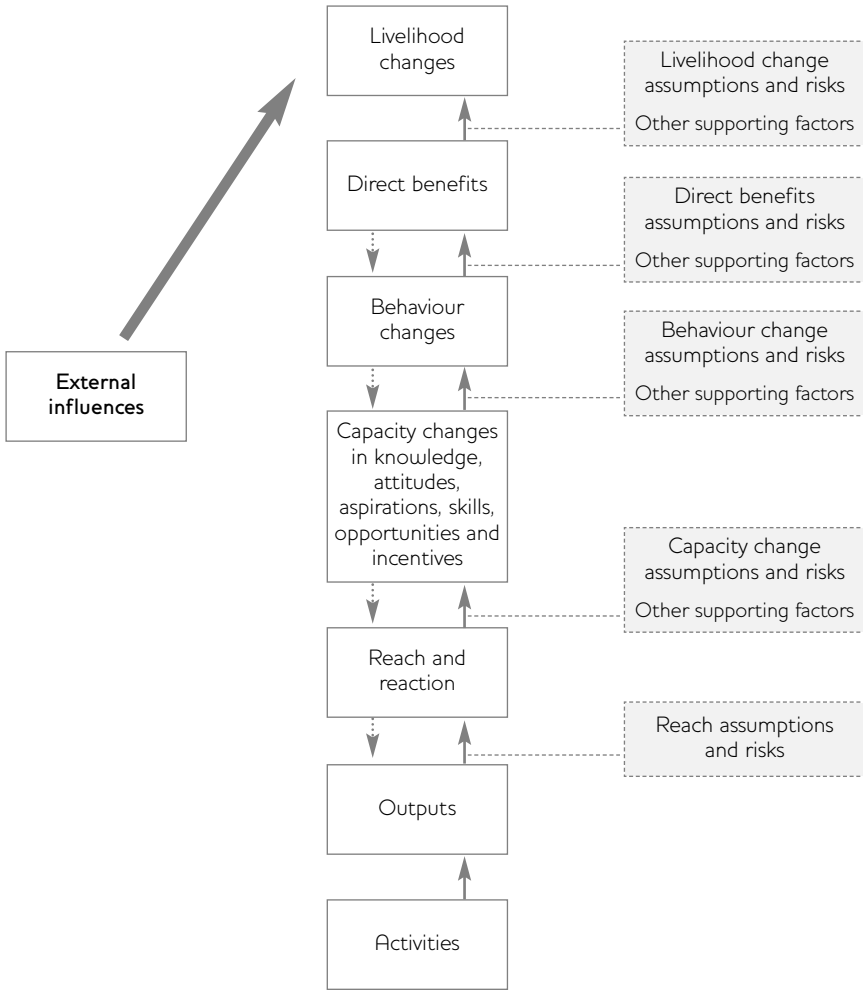
While it is common to adopt a counterfactual view on causality (Gertler *et al.* 2011; HM Treasury 2011; Leeuw and Vaessen 2009; Duflo, Glennerster and Kremer 2008), as discussed by Befani (2012) in Stern *et al.* (2012), and later by Mayne (2012b), there are other approaches to considering causality, in particular regularity approaches, configurational approaches and generative approaches (see Table 1).

Counterfactual analysis is often used to 'associate' or attribute a cause to a given effect, but it does not explain how the effect came about. It merely recognises the existence of the causal linkage, without describing its 'inner workings' (or without 'opening the black box'). Other approaches based on 'regularity' are also limited to associating cause and effect, although – rather than using counterfactual logic – they are based on the frequency of association. Their main weakness is temporal precedence: even when proving a strong association between two factors, it might remain unclear which factor is the cause and which factor is the effect (Brady 2002). More generally, approaches aimed at mere association of cause and effect analyse the relation between one single cause and one single effect at a time (Mill 1843), which in a world of complex interventions aimed at achieving complex outcomes is often inadequate.

Causal analysis can also aim to describe a causal linkage in depth, describing *how* the effect was produced. An effect is sometimes produced by a complex combination of causes and the point of research is then understanding what these causes are, what their role in the combination (or causal 'package') is, and eventually gaining insight on how these factors are combined in the causal mechanism (Glennan 1996). Configurational approaches attempt a first, parsimonious description of the 'ingredients' of the causal mechanism, by identifying configurations of conditions (as opposed to single independent causes) that are associated with the outcome (in the form of conjunctions and disjunctions, aka logical intersections and unions) (Rihoux and Ragin 2009; Ragin 1989). However, they do not go as far as generative approaches in describing the fine-grained characteristics of the causal mechanism.

In evaluation, the most famous application of generative causation is realist evaluation (RE) (Pawson 2013, 2006; Pawson and Tilley 1997). The basic analytical unit of RE is the Context-Mechanism-Outcome (CMO) configuration, which explains an outcome by way of how an individual's (or group's) thinking, acting or decision-making is influenced by contextual resources (e.g. provided by a programme), which can be socioeconomical, human capital, legal frameworks, cultural, etc. In realist evaluation, the mechanism explains how a given resource led to behavioural change. In other theory-based

Figure 1 Basic generic intervention ToC



Source Authors' own.

approaches (including CA, see Section 1.1.1) a more generic theory of change (not necessarily in CMO form) is used to explain a sequence of causal steps leading from the intervention activities to the behavioural change of the stakeholders, and eventually to the intended outcome (Weiss 1997a, 1997b, 1995).

1.1.1 Generative causality in contribution analysis

In evaluation, theories of change (ToC) are used to model how an intervention is expected to bring or has brought about intended changes (Weiss 1997a, 1997b, 1995). These are models of generative causality, showing the steps that occur between some deliberate actions of an intervention and subsequent observed changes, and the assumptions needed for the steps to occur. Figure 1 illustrates a basic ToC for an intervention, showing generic steps between the actions of the intervention and the subsequent changes, along with the assumptions for each causal link and associated risks. The *assumptions* are the events and conditions that need to occur, according to the ToC, if the causal link is to be realised. *Risks* are the inverse of assumptions; if they occur the link does not hold. Also noted are other influencing factors,[1] events and conditions that could or might have influenced the causal link in question positively or negatively, other than the assumptions and risks. While appearing linear, the ToC allows for numerous possible feedback loops.

For most interventions of interest to work, a number of actions, events and/or conditions in addition to those of the intervention are needed

Box 1 Key steps in contribution analysis

Step 1 Set out the cause–effect issue to be addressed

- Acknowledge the causal problem for the intervention in question
- Scope the problem: determine the specific causal question being addressed; determine the level of confidence needed in answering the question
- Explore the nature and extent of the contribution expected from the intervention
- Determine the other key factors that might influence the realisation of the results
- Assess the plausibility of the expected contribution given the intervention size and reach.

Step 2 Develop the postulated ToC and risks to it, including other influencing factors

- From intervention documents, interviews and relevant prior research, develop the postulated ToC of the intervention, including identifying the assumptions and risks for the causal links in the ToC
- Identify the roles other key influencing factors may play in the ToC
- Determine how contested the postulated ToC is to better understand the strength of evidence needed.

Step 3 Gather the existing evidence on the ToC

- Gather the evidence that exists from previous measurement, past evaluations and relevant research to assess the likelihood: (1) of the expected results, assumptions and risk being realised; (2) of each of the causal links in the results chain occurring; and (3) of the other influencing factors making a significant difference.

Step 4 Assemble and assess the contribution claim, and challenges to it

- Set out the contribution 'story' on the likelihood that the intervention 'worked': the causal claim based on the analysis of logic and evidence so far
- Assess the strengths and weaknesses in the postulated ToC in light of the available evidence, and the relevance of the other influencing factors – which links seem reasonable and which look weak and need more evidence
- If needed, refine or update the ToC.

Step 5 Gather new evidence from the implementation of the intervention

- With a focus on the identified weaknesses, gather data on the ToC results that occurred, the assumptions and risks associated with the causal links and the other identified influencing factors.

Step 6 Revise and strengthen the contribution story

- Build a more credible contribution claim based on the new data gathered
- Reassess its strengths and weaknesses, i.e. the extent to which the results, assumptions/risks and other influencing factors occurred
- Conclude on the strength of the ToC and the role played by other influencing factors, and hence on the contribution claim
- If the evidence is still weak, revisit Step 5.

Source Adapted from Mayne (2012b, 2011).

– the supporting factors or assumptions in a ToC. That is, the intervention is unlikely to be the sole cause of a subsequent change, rather it is a contributory cause. What is expected is that the 'causal package' of efforts by the intervention plus other supporting factors together will be sufficient to bring about a desired change (or at least likely sufficient). And further, it is expected that the actions of the intervention are an essential part of the causal package, in which case the intervention can claim to have *made a difference* – not by itself, but within the causal package[2] – and as such is making a causal claim (Mayne 2012a).

This is not the place for a detailed discussion of contributory causes which are neither necessary nor sufficient to bring about an outcome (see Befani 2013 and Mackie 1974 for a discussion of INUS (Insufficient but Necessary part of an

Box 2 Conditions needed to infer causality in CA

1 Plausibility. The intervention is based on a reasoned ToC: the chain of results and the assumptions behind why the intervention is expected to work are plausible, sound, informed by existing research and literature, and supported by key stakeholders.

2 Fidelity. The activities of the intervention were implemented as outlined in the ToC.

3 A verified ToC. The ToC is verified by evidence: the chain of expected results occurred, and the causal assumptions held.

4 Accounting for other influencing factors. Context and other factors influencing the intervention are assessed and are either shown not to have made a significant contribution or, if they did, their relative contribution is recognised and included in the ToC, as part of a larger causal package that the ToC captures as faithfully as possible.

Unnecessary but Sufficient conjunction) and SUIN (Sufficient but Unnecessary part of an Insufficient but Necessary disjunction) causality). However, the above distinction between different causal frameworks helps set the stage for our discussion, and to some extent comparison, of CA and PT. The type of causality we are talking about in both CA and PT is in fact multiple (multiple factors can be responsible for the outcome) and conjunctural (factors combine in complex ways to produce outcomes), so that it is impossible to draw an inference between one single factor and an outcome. It is causal packages and configurations of factors that produce outcomes.

1.2 Contribution analysis

CA (Mayne 2012b, 2011) is based on a ToC for the intervention being examined in detail. Depending on the situation, the ToC may be based on the expectations of the funders, the understandings of those managing the intervention, the experiences of the beneficiaries and/or prior research and evaluation findings. The ToC may be developed during planning for the intervention – the ideal approach – and then revised as implementation occurs, or be built retrospectively at the time of an evaluation. Good practice is to make use of prior research on similar interventions as much as possible. The analysis undertaken examines and tests the ToC against logic, the data available from results observed and the various assumptions behind the ToC, and examines other influencing factors. The analysis either confirms the postulated ToC or suggests revisions to it where the reality appears otherwise. The overall aim is to reduce uncertainty about the contribution an intervention is making to observed results through an increased understanding of why results did or did not occur, and the roles played by the intervention and other influencing factors.

Six key steps in undertaking a CA are set out in Box 1, adapted and expanded from Mayne (2012b, 2011). These steps are often part of an iterative approach to building the argument for claiming that the intervention made a contribution and exploring why or why not.

CA argues that if one can verify or confirm a ToC with empirical evidence – that is, verify that the steps and assumptions in the intervention ToC were realised in practice, and account for other major influencing factors – then it is reasonable to conclude that the intervention in question has made a difference, i.e. was a contributory cause for the outcome. The ToC provides the framework for the argument that the intervention is making a difference, and the analysis identifies weaknesses in the argument and hence where evidence for strengthening such claims is most needed.

Causality is inferred from the conditions and evidence outlined in Box 2.

In the end, conclusions are reached – a *contribution claim* about whether the intervention made a difference, and on how the results were realised.

What does a contribution claim look like? The result of a CA is rarely definitive proof. Causality in relation to socioeconomic interventions is usually of the probabilistic form: that the intervention is most likely to have made a

difference. However, unlike statistical approaches based on large samples, CA builds on different sources of evidence to make an argument from which it is reasonable to conclude with confidence that the intervention has made a contribution, explaining why it did. It builds a compelling case – a warrant – about the contribution being made.

1.3 Process tracing

Process tracing (PT) is a method for within-case analysis[3] that was originally formalised by Bennett (Bennett 2010; George and Bennett 2005), further developed by Collier (2011) and eventually advanced into a textbook by Beach and Pedersen (2012). In PT, a causal mechanism that is believed to explain the outcome is theorised in the form of a number of interlocked components that are all necessary for the causal mechanism to exist (Beach and Pedersen 2012).[4] These components represent entities (for example, actors and institutions) that engage in some kind of activity or display a particular behaviour; these (necessary) components can be seen as a sequence of linked intermediate effects that explain how actions have led to certain outcomes. In both CA and PT approaches, the aim is to look for evidence that increases our confidence in the existence or non-existence of the causal mechanism or the ToC, by increasing our confidence in the existence of its component parts. Below are some general characteristics and principles of PT, which in some cases we will compare with CA.

Firstly, the basic empirical units in PT are so-called 'causal process observations' (CPOs), which differ from 'data set observations' (DSOs) mostly used in quantitative analysis (Collier, Brady and Seawright 2010). Applying PT usually entails observing a causal process that has occurred over time and CPOs are accounts of these observations, while DSOs are matrices displaying the characteristics (variable values) of a sample of cases at a given time. This is in line with analysing a ToC in CA, where causal steps follow a temporal sequence, even though the latter might not be linear and there might be causal loops at work between steps.

In PT, observations[5] are not strong or weak evidence *per se*; for example, observing one step of a process in isolation is usually not very informative. It becomes so or not, depending on what other observations have been made in previous steps of the process or in other parallel processes. In other words, 'evidence' is a combination of observations and other contextual factors such as previous knowledge, timing, the way in which the facts emerge, and so on. PT evidence always results from some combination or accumulation of empirical observations and other contextual information; it follows that the same observation made in different contexts can have very different levels of inferential leverage. This is intuitive if we think about how evidence is dealt with in law courts: it is combined with the circumstances of discovery, the crime, the suspect's motives and past history, and so on.

A general difference with statistical methods is that, in PT, the ability of single observations to act as evidence, evaluated in a context-sensitive manner, is much more important than the number of pieces of evidence collected (Bennett 2008).[6] The quantity of observations can affect the strength of evidence if they are independent, but the basic principle underlying the quality of evidence is not sample size, but rather the probability of observing given pieces of data. These probabilities can be general, like the average probability of observing a given piece of evidence P(E), or conditional, like the probabilities:

- of observing that same piece of evidence under the assumption that the causal mechanism holds, or P(E|CM)
- of observing that same piece of evidence under the alternative assumption that the causal mechanism does not hold, or P(E|~CM).

If observations made seem to confirm the causal mechanism, some questions to ask are: 'Do these observations support other causal mechanisms? Could these observations have been made if the change was being caused by other causal mechanisms? If so, how likely would these observations have been? Do these observations seem to be unique to this causal mechanism?'.

Other important probabilities in process tracing are: (a) the prior probability of the causal mechanism being triggered, P(CM), which expresses our confidence that the mechanism holds, on the basis of previous knowledge from different sources and prior to conducting data

Box 3 Criteria to assess the strength of evidence in Bayesian analysis

The strength of evidence E for a causal mechanism CM is proportional to the distance between the probability of E under the causal mechanism [P(E|CM)] and the average probability of evidence E [P(E)], and to the distance between the probability of E under the causal mechanism [P(E|CM)], and the probability of E under all alternatives to the causal mechanism P(E|~CM).

collection (related to the principle of plausibility in CA); and (b) the probability that the causal mechanism holds after data collection, when a given piece of evidence has been observed, P(CM|E). The latter is what we try to maximise through data collection: we aim to design data collection so as to maximise its power to change our prior, 'theoretical' confidence. If P(CM|E) is similar to P(CM) it means that our evidence is weak: the bigger the difference between the two probabilities, the stronger evidence E is.

As researchers we often use these probabilities unknowingly, almost unconsciously: for example, we know that the presence of vested interests in the success of a particular project will make judgement about that project by an individual with such interests biased/unreliable; his/her positive judgement will be weak evidence that the project was actually successful. In formulas, this is because the prior probability of that individual evaluating the project positively [P(E)] is in any case (at the start, independently of how the project actually goes) high. It is true that, if the project is successful, s/he would have likely given a positive judgement [P(E|CM) is high]; but because s/he has stakes in the project and other incentives, she might give the same judgement also for other reasons, if the project is not successful [P(E|~CM) is also high].[7] The difference between these two probabilities is low, hence the weakness of evidence, E.

Bayes' theorem shows us why P(E) being high is not good for the quality of evidence. The Bayes formula indicates how our initial confidence P(CM) about a causal mechanism, CM, is changed by the evidence collected (E), becoming P(CM|E). It can be illustrated in two different but equivalent ways. The first:

$$P(CM|E) = P(CM)*P(E|CM)/P(E)$$

shows that an observation, E,[8] strengthens our initial confidence in the causal mechanism [which is represented by P(CM)] if the prior probability of the observation P(E) is low, because P(E) stands in the denominator of the ratio by which P(CM) is multiplied. In particular, our confidence is strengthened by the observation E if its probability under the theory of change P(E|CM) is higher than its general, 'average', probability P(E).[9] We can also see that the higher the difference between the expectation of the evidence under the theory holding P(E|CM) and the general expectation of the evidence P(E) the more our post-observation confidence in the theory of change P(CM|E) is strengthened.[10]

The second formula illustrating Bayes' theorem[11] shows that an observation E increases our initial confidence in the causal mechanism P(CM) if the probability of making that observation if CM is realised [P(E|CM)] is greater than the probability of making it if CM is not realised (i.e. under alternative assumptions) P(E|~CM).[12]

1.3.1 Using evidence in PT

The Bayes formula as illustrated above implies that we use evidence in order to increase or decrease our confidence P(CM) in the existence of a causal mechanism, CM, and provides criteria to judge the power that evidence has to change our pre-observation confidence. We summarise these criteria in Box 3, while we address their practical implications later.

The above criteria are related to probabilities and, therefore, to some extent are quantitative. Luckily PT also provides qualitative criteria to judge the quality of evidence: certainty and uniqueness. Usually we cannot be absolutely certain that a causal mechanism holds; however, we can be certain that some mechanisms do not hold. Certainty thus refers to the ability that some tests have to rule out causal hypotheses, in other words it refers to the disconfirmatory power of tests. These are called hoop tests because the causal candidate needs to 'pass through the hoop' if it is to be retained as a possible cause (Beach and Pedersen 2012; Collier 2011; Bennett 2010; Van Evera 1997).

Table 2 **Relation between types of observations, types of evidence, Van Evera's tests and the Bayes formula**

	Description: an observation (or a set of observations) that is:	Relation to Van Evera's tests	Relation to the Bayes formula P(E|CM)	P(E|~CM) and P(E)
Disconfirmatory evidence	Unlikely under the causal mechanism (CM)	Fails hoop test	Low	–
Confirmatory evidence	Unlikely under any alternative to the causal mechanism (~CM)	Passes smoking gun test	–	Low
Both	Likely under the causal mechanism (CM); unlikely under any alternative (~CM)	Doubly-decisive test	High	Low
Neither	Likely under the causal mechanism (CM); likely under alternatives (~CM)	Straw-in-the-wind test	High	High

Source Authors' own.

Uniqueness is less strong as it refers to the confirmatory power of tests. Unique tests (called smoking gun tests) can confirm that a causal mechanism is indeed at work on the basis of the 'signature' traces it leaves (which are deemed to be unique to that mechanism and practically impossible to have been left by other mechanisms). If a unique test confirms the presence of a mechanism, this does not necessarily mean that the outcome was exclusively produced by that mechanism: it might just mean that the latter has a causal role, that it is a contributory cause. Causality is often multiple and smoking gun tests might not provide information on other mechanisms that might be at work in parallel to the one we are confirming. They might just confirm that a particular mechanism has a role. We will see in Section 1.3.2 that there are different types of unique tests, and not all of them compare mutually-exclusive hypotheses (Rohlfing 2013).

The typology of PT tests includes two additional types: the straw-in-the-wind and the doubly-decisive. We will not address the former as it is neither confirmatory nor disconfirmatory,[13] while the latter is both, being simultaneously characterised by certainty and uniqueness, and being able to both confirm a causal mechanism and reject all of its alternatives.

There is a link between the probabilistic formulation addressed above and the qualitative tests. The hoop or disconfirmatory test is a piece of evidence that we expect to observe if the causal mechanism holds: in other words the probability of observing E under CM is high, which means the numerator of the Bayes formula P(E|CM) is high. The smoking gun or confirmatory test, on the other hand, is a piece of evidence that is unlikely under any alternative to the causal mechanism [P(E|~CM) is low] and thus its average probability P(E) is low, too. As seen above, those quantities can be found in the denominator of the Bayes formula showing how our pre-observation confidence in the causal mechanism P(CM) changes after data collection [into P(CM|E)].

In practice, these tests guide the search for evidence. Disconfirmatory tests prompt the question 'What do we expect to observe if the causal mechanism is realised?', while confirmatory tests ask 'What observations could only be made if the causal mechanism holds, and could not be made if it does not? What observations are unique to the causal mechanism?'.

In other words, confirmatory evidence is an observation or a set of observations that could not have been made, or is extremely unlikely, under alternatives to the causal mechanism, while disconfirmatory evidence is an observation, or a set of observations, that could not have been made if the causal mechanism under test were realised (see Table 2).[14]

1.3.2 The different implications of confirmatory evidence

A specific piece of evidence might confirm a causal mechanism, but other observations might well support the parallel existence of others. So

confirming the CM does not automatically invalidate other explanations, even though sometimes it could. This means that not all alternatives are 'rival' in the proper sense.

In an attempt to understand the differences among confirmatory tests and their implications, a typology of ten different types of confirmatory tests, all having a working hypothesis and an alternative, has been proposed (Rohlfing 2013). These tests differ on the characteristics of the cause and the effect. Namely, the two alternatives being compared might have the same cause and the same outcome, or mutually-exclusive causes or outcomes, or non-mutually-exclusive causes or outcomes. Two such tests are of particular interest here. We introduce them briefly below and then use them in the application section.

The first type of confirmatory test attempts at answering the question: '(On the basis of the collected data) did the intervention (likely) contribute to the outcome? Or did factor F likely contribute?'. In symbols, the working assumption is represented by I -> O, while the alternative by F -> O.[15] In this case confirming the causal relevance of the intervention does not provide any information on the relevance of the other factor(s). If we want to learn about the latter, we need to repeat the same test, using factor F as the working assumption. This way we might answer the question: Were there other contributing factors, in addition to or in combination with the intervention?

The second type of test is applied to our strongest hypothesis, usually a complex, 'heavy' ToC including several factors and mechanisms that have been shown to have contributed to the intervention separately. The question is: 'Is this ToC the only (most) plausible explanation for the outcome?'. In symbols, the working hypothesis is ToC -> O, while the alternative is ~ToC -> O. This is a doubly-decisive test because, unlike the previous test, confirming the ToC automatically invalidates any other causal mechanism. We will use these tests in Section 2, where we apply some of the concepts introduced so far (certainty, uniqueness and the various probabilities) in examining the case of an intervention providing training to teachers in an attempt to increase the school performance of girls.

2 Carrying out contribution analysis applying the tests and principles of process tracing

The case presented here is not a full, proper application of the procedure, but rather an attempt to imagine how PT can be applied to a real-life evaluation that has been conducted in the past using CA. It is not meant to represent an ideal case for PT, but more like a 'test case' for a 'proof-of-concept' argument. It will perhaps not provide operational guidelines, but it does sketch the contours of a procedure, which we call the CA-PT procedure, and which we hope deepens the discussion on how PT and CA can be combined to strengthen each other.

One reason why we are trying to combine CA with a method is that CA is an approach, and does not spell out detailed steps to follow in data collection or discusses explicitly the types and strength of evidence used. Applying PT to a CA can thus allow us to use the logic of CA as overarching guidance, while at the same time ask specific questions related to data collection, such as: 'What kind of evidence is (mostly) necessary and/or (mostly) sufficient to confirm/disconfirm a causal explanation?'. In other words, PT provides CA with indications on what evidence to look for and with criteria to judge the strength of that evidence, complementing the CA steps outlined in Box 1, in particular steps five and six.

In this section we try to answer the PT question both in relation to the whole ToC as a 'heavy' mechanism including and combining many causal factors, and in relation to its single steps/components, starting from the latter. We will see that these two objects require two different types of PT tests.

We present a procedure to carry out the last two steps of CA, inspired by PT, which is made of three broad steps[16] that we will call PT steps, and which are tests of:

1 the intervention main mechanism (PT1);
2 other causal mechanisms external to the intervention (PT2);[17] and
3 the comprehensive ToC including the intervention and the external factors (PT3).

PT1 and PT2 entail carrying out hoop tests and smoking gun tests for the causal mechanism under analysis: the intervention main

Table 3 **Relation between PT steps and PT test types**

		PT test types		
		Disconfirmatory phase (hoop test, maximising certainty)	**Confirmatory phase (smoking gun test, maximising uniqueness)**	**Doubly-decisive test (both certainty and uniqueness)**
PT steps	PT1: testing the intervention	Yes	Yes	No
	PT2: testing other potentially contributing factors	Yes	Yes	No
	PT3: testing the complex ToC as a whole	No	No	Yes

Source Authors' own.

mechanism (PT1) and the other causal mechanisms external to the intervention, to be tested one by one (PT2). Finally, if the evaluator feels confident enough that s/he has good evidence on all relevant causal factors, s/he can try to test the whole ToC in a doubly-decisive test (PT3). Straw-in-the-wind tests are not considered here as they neither confirm nor disconfirm the hypothesis (see Table 3).

2.1 Setting up the case

We use a hypothetical example of an intervention that has been in place for several years aimed at improving girls' education outcomes.

2.1.1 The problem, as seen at the time the intervention started

In a region, education outcomes for girls are low. Girls' education is not seen as a priority; attendance is low and drop-out rates are high. Based on some interviews with students, households and teachers, a likely key issue is that many teachers themselves do not see girls' education as important and do not have the skills to provide a gender-sensitive approach to education. Physical access to schools *per se* does not appear to be a problem. Besides lack of teacher preparation, other factors that likely contribute to low educational outcomes are the availability of work on the market for girls (in a nearby garment factory), the expectation that it is a priority for girls to help the older women with household chores, discomfort felt by girls with school accommodation, and the poor availability of textbooks. The organisation responsible for implementing the intervention has limited resources and can only focus on improving teacher preparation; however, during the implementation of the programme, other interventions attempt to tackle some of the other issues.

2.1.2 The intervention

Based on this analysis, the intervention has been set up to provide special gender-sensitive training to teachers, stressing the importance of education for girls. It provides training to teachers to raise their awareness of the special needs and situation of girls in school, and addresses teachers' attitudes towards education for girls. It also provides them with ways and means to adopt a more gender-sensitive approach in classrooms, by ensuring that girls are not discriminated in their opportunities to learn.

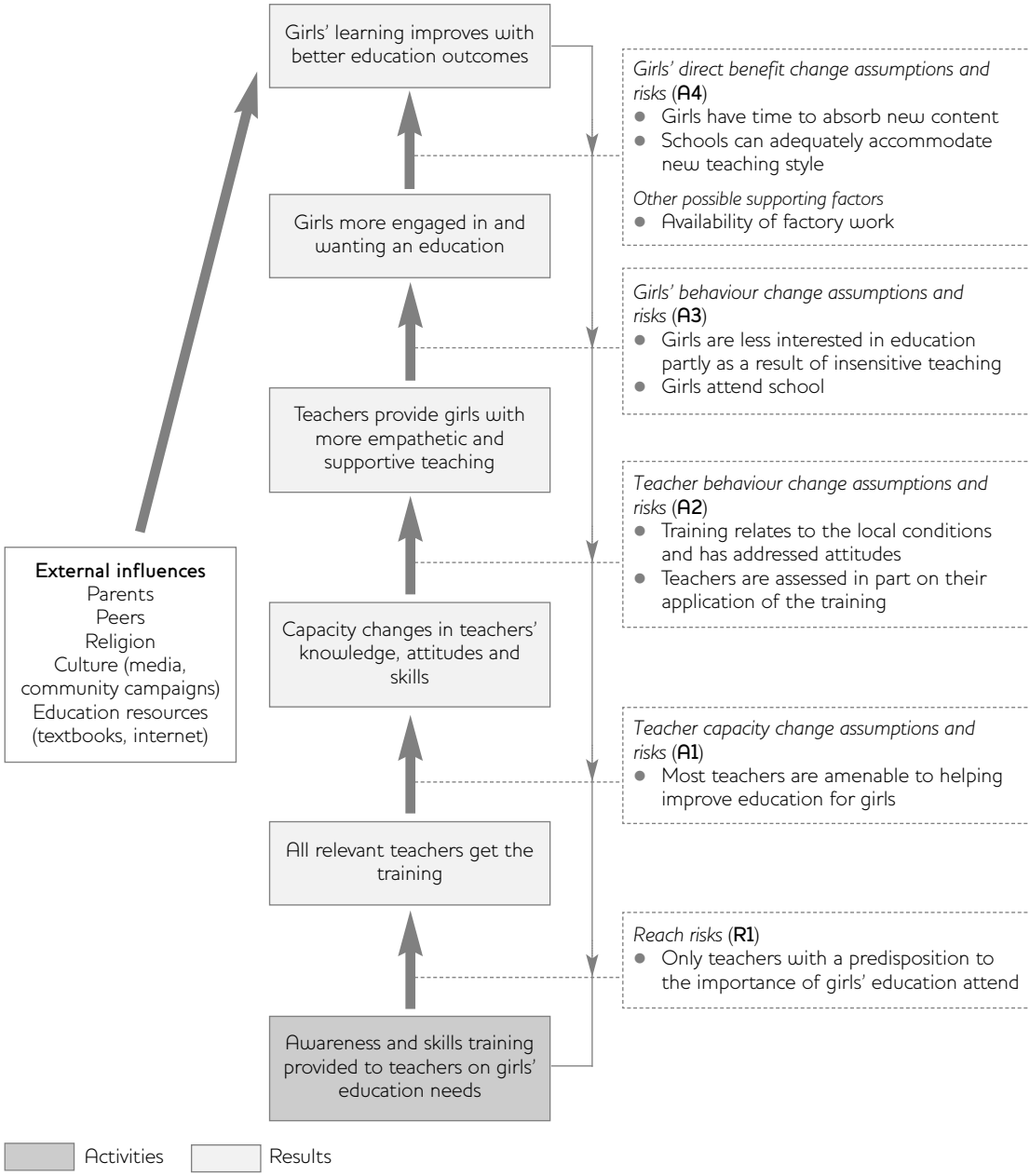
2.1.3 The overall ToC

The intervention is based on the idea that *if* teachers provide a more supportive and gender-sensitive teaching in the schools, *then* girls will be more actively engaged in studying and wanting an education, which will lead to better education outcomes for girls (which can be measured, for example, with test scores or graduation rates).

2.1.4 Prior evidence for the intervention

The quality of teaching has been shown to be a significant factor in better education outcomes (OECD 2005). However, there is poor evidence that sensitising teachers to the specific problems faced by girls will result in better educational outcomes for girls, partly because this type of intervention has been rarely implemented (and evaluated).

Figure 2 ToC for enhancing education outcomes for girls



Source Authors' own.

2.1.5 Main evaluation findings

After two years, an evaluation is undertaken and it is found that indeed education outcomes for girls have improved: the average test scores are higher.

The main evaluation questions then would be:

- Can the improved outcomes be 'claimed by' the intervention?
- How was the improved school performance of girls brought about?

When we apply PT we assume that CA has been completed up to and including step four, and that data collection on the implementation is about to start. We propose to show how the data collection process to answer the above questions could be designed using PT.

2.1.6 A ToC for the intervention

Contribution analysis starts with a ToC for the intervention. Ideally this is one developed during the planning stage of the intervention and revised over time as more data and understanding are acquired. If this was not done, then a ToC is developed at the outset of the evaluation and perhaps revised at the end. Developing a ToC was briefly discussed earlier.

A ToC for the intervention is shown in Figure 2. The assumptions listed are events and conditions necessary for the causal link in question to occur. From the assumptions listed in the ToC (and noted in setting out the problem) it is clear that considerably more than teacher training is likely needed to make an impact on the girls' education. These other supporting factors such as the ability to accommodate girls, a supportive home environment, and teachers' performance being assessed in part on how they have applied the training – to name only a few – along with the teacher training provided comprise the intervention 'causal package'. It is expected that over the course of the evaluation, the ToC will be revised to better reflect the context of the intervention and the causal factors at work.

Figure 2 lists some other possible influencing factors. In the course of the evaluation, students, teachers, parents, community leaders, school administrators, etc. would be asked to identify any changes that had occurred during the past two years that they feel might have influenced the good educational outcomes achieved and why they think so. This would provide a rich source of factors that might have had an influence.

*2.2 **The analysis***

At this stage CA has been completed up to and including step four,[18] and data collection needs to be designed and conducted in order to carry out step five, and eventually step six (perhaps iteratively). There are a number of plausible causal factors in the ToC, and the PT-inspired procedure starts by testing them. The first phase (called the disconfirmatory or ruling out phase) is aimed at ruling out some of these factors (including possibly the contribution of the intervention) relatively quickly, with the confirmation work to be conducted later. This is the hoop test phase, aimed at maximising certainty, because hoop tests are used for disconfirming the hypotheses about the relevance of the causal factors, and lays out the foundation for PT1 and PT2 because it is conducted on all causal factors considered, including the intervention.

2.2.1 Maximising certainty: ruling out possible causal factors through hoop tests

This phase is about clarifying your ideas about what you expect to observe if a given causal factor contributed to the outcome and then testing your expectations on reality. You can measure the strength of your expectations with the probability of making a given set of observations if the causal mechanism for the factor holds, which in symbols is P(E|CM) (it sits at the numerator in the Bayes formula). Consider two causal mechanisms:

Causal mechanism A (CMA): the training contributed to increased school performance of girls. If the training had an impact, the evaluator has at least two expectations for the moment (we will call these both EA for evidence of causal mechanism A): (1) that the training has been provided to a considerable number of teachers, and (2) that the girls acknowledge the contribution of teachers by providing a judgement of their performance between neutral and positive. The latter is not sufficient to claim impact, as it is thought that girls would be afraid to say anything negative about teachers, but it is thought of as necessary because if there are widespread complaints about teacher performance from the girls then it becomes unlikely that outcomes improved because of anything the teachers did. Similarly, the extent of training provided is not sufficient to claim impact but it is necessary as the outcomes are widespread. In sum, failing to observe any of the above two facts greatly reduces our confidence that the intervention had an impact, and under normal circumstances, eliminates the hypothesis, shifting the evaluator's attention to other factors. This is because the probability of observing the two EA factors under the assumption that causal mechanism A holds is high [P(EA|CMA) is high]. With regard to Table 2, we would be in a situation where the observations made (~EA) would be unlikely under the hypothesis CMA, and the hoop test fails (first row).

Causal mechanism B (CMB): the closing of the local garment factory meant that girls were no longer sent to work there and had more time to study, hence the increase in educational outcomes. If this explanation is valid, we would expect to observe: the closing

of the local factory, evidence that it was employing many girls, and that these girls have actually been granted free time instead of being employed elsewhere or in assisting with household chores. We call this EB for evidence of causal mechanism B: if the causal mechanism is realised, the probability of observing EB is high [P(EB|CMB) is high]. That is why failing to observe EB greatly reduces our confidence in causal mechanism B, to the point that we lose interest in the explanation.

If all sets of potential observations EA and EB are actually observed, both causal mechanisms pass the respective hoop tests, which means they remain causal candidates for the outcome. Usually at this stage we still do not have strong evidence that those causal mechanisms were actually at work, which we look for in the next phase. Conversely, say if EB is not observed because no factory was closed, CMB fails the hoop test and only CMA remains as a causal candidate for the outcome.

Similar arguments would be made for each identified causal factor. The TOC could be then revised, putting more emphasis on those causal factors that survived the hoop test, and downplaying the role of, or removing, those that failed it. If the intervention mechanism fails the hoop test, the evaluation task could potentially end here, unless the evaluation questions are changed to increase the focus on other factors.

2.2.2 *Maximising uniqueness: verifying the detailed steps of a sophisticated causal mechanism*

In this phase the causal mechanisms that have survived the hoop tests are administered confirmatory or smoking gun tests, aimed at finding unique evidence that practically confirms or greatly increases our confidence in their existence. Unique evidence for a causal mechanism is a set of observations considered (extremely) unlikely under any other causal mechanism. This increases our confidence in the hypothesis because unless that hypothesis was confirmed, making that specific set of observations would have been (extremely) unlikely. In Bayesian terms, the probability of making those observations P(E) is in general (extremely) low, or at least (extremely) low under alternatives to the causal mechanisms P(E|~CM) (these probabilities sit at the denominator in the Bayes formula).

Constructing confirmatory/smoking gun tests is more complicated than constructing disconfirmatory/hoop tests. In order to minimise the probability of observing evidence that we hope to be confirmatory, we need to formulate the ToC in a way that is as unique, specific and detailed as possible, with a high number of clearly specified steps, a high level of detail on the inner workings of the causal mechanism, including resources; incentives; preferences; thinking; actions and interactions of all the stakeholders involved in the causal mechanism. Building the mechanism as a concatenation of several normally independent events will ensure that the probability of observing the entire set of events P(E) is minimised, following the probability law assigning to the combination of n events x1 * x2 * … xn the product of the probabilities of the single events P(x1) * P(x2) * … * P(xn). P(E) becomes increasingly lower (and the strength of the evidence when E is observed increasingly higher) as more and more (independent) components are added to the causal mechanism.

The extent to which this can practically be done will vary by situation. Further, in cases such as we are exploring where there are very likely to be multiple causal factors at play, the uniqueness of the effects of the activities of the intervention – the training in our case – may only be apparent for proximate effects (see Ton *et al*., this *IDS Bulletin*). Further along the impact pathway, other causal factors might explain the causal mechanisms at work.

We start from wanting to confirm the causal mechanism explaining the contribution of the intervention, without intending to automatically weaken other explanations. As we collect more evidence on the relevance of several coexisting causal factors or mechanisms (as suggested in CA condition four, Box 2), we then become more ambitious and build an increasingly complex ToC interconnecting all the factors and mechanisms shown to be relevant. At some point we might feel confident enough to try a doubly-decisive test.

Table 4 shows the relation between the PT steps as defined above, CA steps and CA conditions (see Box 1 and Box 2). Steps one to four are not addressed in detail here but they all collectively contribute to our prior confidence that the theory or its causal mechanisms hold P(CM).

Table 4 **Relation among PT tests and CA steps and conditions in the combined CA-PT procedure**

		CA conditions (Box 2)	CA steps (Box 1)
PT steps	*PT0: building confidence on the causal mechanism on the basis of previous knowledge pre-data collection*	1 and 2	1, 2, 3 and 4
	PT1: testing the intervention	3	5 and 6
	PT2: testing other potentially contributing factors	4	5 and 6
	PT3: testing the complex ToC as a whole	–	5 and 6

Source Authors' own.

2.2.2.1 Assessing the contribution of the intervention to improvement of girls' school performance (PT1)

Part of step one was carrying out the hoop test on this causal mechanism, i.e. the intervention, and if we are at this stage it means that the hypothesis survived it. Another causal mechanism also survived the disconfirmatory phase: the one about the factory closing down. As our knowledge about the relevant causal factors at this stage is still relatively poor, we start from a test where confirmation of one causal mechanism does not automatically eliminate others (i.e. where causal mechanisms are not mutually-exclusive, Rohlfing 2013). The relevant general question is the following:

Did the intervention contribute to the outcome (i.e. the working hypothesis is I -> O)?

- In this test, confirming the causal relevance of the intervention does not provide any information on the relevance of other factors, which will need separate testing using the same uniqueness principle. So if the hypothesis that the intervention contributed is confirmed, this does not mean that the closing down of the factory did not also contribute.

The key specific question here is: What observations can only be made if the intervention has indeed made a contribution?

Identifying a unique set of observations is extremely useful because, when made, they practically prove impact. In order to identify this unique set of observations that we wish to make, we need a detailed and sophisticated version of the causal mechanism explaining the contribution of the intervention, and then – in line with the third condition of contribution analysis – hope to confirm/observe all of its components.

The sophisticated mechanism we are trying to map out identifies a specific process linking what the teachers were taught, with what they implemented in their teaching practice, with the girls' behavioural change, and finally the improved educational outcomes. Figure 2 shows this impact pathway. These linkages might be tested by answering the following evaluation sub-questions:[19]

- Was the training well received by teachers?
- Was the teaching offered to girls improved along the lines of the training that was delivered?
- How likely would these improvements have been without the training?
- Can girls relate in specific ways their increased interest in learning to the teaching they are now getting?
- Is the better teaching among the factors that stakeholders (girls, parents, administrators, community leaders) point to when explaining the improved educational outcomes for girls, particularly when prompted in a way that reduces the probability of them mentioning the teaching?

Answering the above questions means verifying the existence of all the steps in the causal mechanism relating the training to improved performance. If positive answers are found to all or most questions, the observed causal

mechanism associated with the training constitutes 'smoking gun' evidence that, at least in part, the improvements in teaching are due to the training, particularly if alternative explanations for each step are unlikely, and the causal mechanism describes what they are and how they came about. This process is similar to that carried out in a CA, but the PT approach suggests a focus on tracing the specific effect of the causal factor as far as possible along the impact pathway.

The uniqueness of this test can be high because once we observe all of this it becomes difficult to think of another explanation (alternative to the intervention) that could be responsible for triggering all the steps in the causal mechanism. We can think of alternative events being responsible for each single step, but the chance of all these alternative events materialising to trigger the sequence instead of the intervention being successful is low, because the events are relatively independent. The training might not have been appreciated at all by the teachers, subsequent teaching in the classrooms might not have reflected the training or teachers were unable to implement the correspondent changes in class, girls might have viewed the changed teaching approaches as condescending or superficial, or the improved teaching might have been well received but not viewed as a relevant factor in the better outcomes. The chances of all these events aligning at the same time, one step after the other, are generally low [P(E) is low]. In particular, they are low if the training did not have an impact and these events are triggered by other causal mechanisms unrelated to the intervention [P(E|~CM) is low]. So the empirical confirmation of all of these links is strong evidence that the mechanism is actually at work.

Recall that this particular step is trying to show that the intervention made a contribution to the outcomes, not that it was the sole cause. The strength of the evidence for the impact pathway in question would provide some basis on which to assess if the intervention was a main 'cause' of the outcomes. These later conclusions would be strengthened by assessing the impact pathways for other causal factors that had passed the hoop test, which we address in the next section.

As the Bayes formula is widely used in science and professional practice, as well as more or less explicit forms of the hoop and smoking gun tests, the above causal inference from intervention to outcomes is thus justified on both scientific and common sense grounds.

2.2.2.2 Assessing the relevance of other factors (PT2)

In the previous stage, we have shown that I -> O, but we still do not know if the intervention is the only factor contributing to the improvements. If we are interested in learning about the role of other factors, we can follow a similar procedure to build smoking gun evidence for the other factors that have survived the hoop test, in our case the closing down of the garment factory. We might not necessarily be interested in this, because if the other factors also had an impact, this does not necessarily take anything away from the relevance of our intervention, unless they are independent from or unrelated to the intervention (see PT3). In other words, unlike in statistical linear models and in estimates of net effects, confirmation that one factor has contributed does not necessarily decrease the chances of others having done the same: the contribution of causal factors assessed with CA or PT is not necessarily additive.

However, even in such a case, assessing the contribution of other factors is important for learning why the intervention worked and for the transferability of lessons learned (in other words, to maximise the external validity of the findings). Some of these factors might have 'prepared the ground', allowing a successful performance of the intervention which would not have happened had they not been there.

Now let us assume that through smoking gun tests we end up confirming causal mechanism CMB about the garment factory. It turns out that the closure did not visibly affect the amount of time the girls spent assisting with household chores, but at the same time, it gave girls a lot more time to study. Girls spent the same time helping other women in the house, but now did not have to go to work in the afternoon and had some more time available. Some of these girls used their time to study, which – judging from detailed accounts of how this time helped performance – increased our confidence that the factory closing contributed to the outcome. The other identified causal factors would be explored in a similar way.

2.2.2.3 Testing the ToC as a whole (PT3)

It might be later discovered that not only the closing of the garment factory contributed to the outcome, but also that the additional available study time was essential to make one specific aspect of the training work. Indeed, among the new activities implemented by the teachers after having received the training, was an additional module addressing the opportunities potentially opened up by education, which aimed to help raise awareness of the importance of school for girls. This module was supposed to act as an incentive for girls to devote their energies to school, raising their awareness that performing well at school could make a difference to their future life. But in order to work (to affect outcomes), the additional module needed additional time to be absorbed, which was indeed available to most girls because the garment factory had closed. That part of the training would not have worked if the other causal mechanism had not taken place: in this case the outcome was produced by a combination or package including the two.

At the same time, some of the factors discovered to be relevant might not be directly needed for the intervention to work like, for example, the discovery of an awareness campaign on the importance of girls' education being rolled out in the area, or the visit of a famous girls' rights activist which received considerable media coverage. Rather than being directly 'supporting' to the intervention, we might realise that these factors trigger relatively independent pathways, more distantly connected to the training (the 'external influences' of Figure 2). The argument would be that the teaching did not need the campaign or the visit to improve, nor to improve outcomes, and these factors might be shown to have improved outcomes through their own unique pathways.

In brief, the point of the analysis in PT3 is to analyse the connections among the relevant factors and determine which ones act together, forming a causal package (as in the intervention + assumptions + risks part of the ToC), and which ones act more independently. While the factors in the causal package trigger one single outcome together, making teasing out which factor contributed to what outcome a futile exercise, the independent pathways trigger outcomes that can be distinguished and thus added up to each other. This means that if one of these outcomes was discovered to be very significant, our confidence in the impact of the intervention would decrease (because, for example, there was a religious reform that changed cultural attitudes about girls going to school, which was credited as the main explanation). Conversely, if the effects from the independent pathways are discovered to be weak, our confidence that the 'intervention + supporting factors' package is to be credited the most would increase.

As we learn more about what causal mechanisms did actually take place (in PT2) and about their relation (or lack thereof) to each other (in PT3), we gradually confirm and/or revise the ToC in Figure 2 with the causal mechanisms that have received empirical support; the conditions (assumptions) that were needed in order for those mechanisms to work according to the evidence; the risks avoided; and with external contributory factors that were unrelated to/independent from the intervention but have received empirical support.

At this point the most difficult question can be asked: 'Was it the intervention mechanism with the help of supporting factors, plus other external independent factors, all represented in the ToC, that best explains the outcome? Or not?'. In symbols:

> Working assumption: ToC -> O; alternative assumption: ~ToC -> O

Compared to PT1 and PT2, PT3 is more logical than practical, and it does not require as much data collection, but rather a synthesis of the insights gained so far, and a comparison of the knowledge acquired for the different factors and causal mechanisms. The reason why we list it as a separate PT step is that it has a doubly-decisive value: if the working assumption is confirmed, the alternative assumption is automatically invalidated and the ToC is declared a clear winner.[20] This happens because all plausible factors have been tested and there simply are not any left.

This final test is very demanding, and in practice it will not always be possible to get to a point where all plausible factors have been thoroughly tested and a doubly-decisive test can be conducted. This should not discourage potential evaluators as the procedure does not require all PT steps to be

necessarily undertaken: in many real-life situations PT1 and part of PT2 will be carried out, with a part of the possibly relevant causal factors not thoroughly tested. This should in any case lead to a modified ToC with stronger empirical support than the initial one, and in which we have higher confidence. We would not be able to conclude that this ToC is the best explanation for the outcomes, but we would be more confident (or less confident, depending on the evidence) that some of the factors tested (including the intervention) contributed to the outcomes.

3 Concluding remarks: from 'assessing impact' to 'assessing our confidence' about impact

This article has shown that CA could be a useful entry point to apply PT in impact evaluation and that, at the same time, PT could strengthen inferences made with CA, using tests and principles largely used in science and professional practice. Where CA says 'verify causal links', PT indicates in more detail how to go about verifying them. The article proposes a combined CA-PT procedure made of the traditional CA steps plus indications about how to carry out steps five and six, which are more specific than those currently offered by CA. The procedure comprises three broad steps, each including some PT tests. The intent was not to describe in great detail how an evaluation of this kind might be undertaken, but to show how the logic of the two approaches can be combined in the application to a single practical case.

We have found a substantial overlap between the conditions and steps of CA and PT. The proposed procedure tests the intervention mechanism first using hoop and smoking gun tests, and goes on to use the same tests on other causal factors, some of which are discovered to be essential for the intervention to work, while others, although affecting the outcome, are unrelated. When enough evidence has been accumulated on the relevant factors and causal mechanisms, including how these are related and work in combinations/packages, a final test can be performed on a 'heavy' version of the ToC, the confirmation of which halts the search for additional explanations.

The almost seamless integration of the two methodological approaches and the way they strengthen each other, CA by making PT more evaluative and PT by relating CA to an established methodological approach, proves that their combined use holds promise in exploring and testing alternatives to counterfactual causal inference in impact evaluations of development interventions.

In particular, the proposed procedure does not 'measure impact' but rather assesses our confidence about impact; and the strength of the evidence collected is measured, in accordance with the Bayes formula, on its ability to change our prior, pre-data collection confidence that the intervention actually had an impact.

Notes

1 We define 'causal factors' as factors that might influence/contribute to the outcomes.
2 In the causation literature this is called an INUS (Insufficient but Necessary component of an Unnecessary but Sufficient package) cause (Mackie 1974).
3 'Within-case analysis' refers to studies of one case, where the case study is defined as a single entity as opposed to case-based studies which can compare multiple cases, or studies observing phenomena over a large population of statistical units.
4 This means that the simultaneous presence of all the connected parts is essential to identify the causal mechanism. If one part is lost, the causal mechanism is not the same and needs to be redefined.
5 CPOs are not the only type of data useful in PT; it is possible to distinguish among at least four types of evidence: sequence, pattern, account and trace (Beach and Pedersen 2012).
6 The logic used to assess the strength of evidence in process tracing is inspired by the Bayesian approach to probability, instead of frequentist probability which informs traditional statistical evidence.
7 One difference with a statistical approach is that, if we relied only on sample size, we would not estimate probabilities of specific individuals or groups giving a positive judgement and then compare these probabilities to the judgement actually given, even in small samples, but simply calculate the average judgement about the project from a large random sample of stakeholders.
8 'E' normally stands for 'evidence'; however, since we are referring to observations in a broad sense, and observations are not

necessarily evidence, we will take 'E' to refer to the more humble concept of 'observation'.

9 The sensitivity of the evidence P(E|M) (also known as the ability to minimise Type II error or false negatives) is high. This is because our post-observation confidence in the theory P(T|E) is obtained as P(T) (our pre-observation confidence) multiplied by a factor, and is thus increased if the factor is >1, while it is decreased if the factor is <1. The factor is a ratio of probabilities, and will thus be >1 if the denominator is smaller than the numerator, and <1 if the denominator is greater than the numerator. It follows that our confidence in the theory is strengthened by the evidence if P(E|T) is greater than P(E). In particular, the greater P(E|T) and the smaller P(E), the more our confidence is strengthened.

10 This is intuitive when we think of how we draw inferences in everyday life. We do not start suspecting that 'something is the case' until we observe (perhaps repeatedly) something we do not observe under 'normal' circumstances, something which is more probable 'under the case' than on average. For example, if you are in a shop and hear the door open you do not suspect that someone you share a house with is opening the door. It is true that, if your partner wanted to open the shop's door, you would hear the door open [P(E|it is your partner) is close to one], but you hear the door open very often while you are in the shop so P(E), the general probability that the door opens while you are inside, is also close to one. Conversely, if you hear your house's door open, the likelihood of it being your partner (represented as [P(it is your partner|E)] is much higher because, while it is as true as in the shop that they would be able to open the door if they wanted to [P(E|it is your partner) is the same], your house door opens a lot more rarely than the shop's, because you need a key to open it and only very few people have it [P(E) is much lower].

11 P(CM|E) = P(CM)/[P(CM) + P(~CM)*P(E|~CM)/P(E|CM)].

12 This is also intuitive: back to the previous example, suppose that P is your partner wanting to open the door, while ~P is any other person wanting the same. When you are in the shop and hear the door open, it could be many people in addition to her/him. If she wanted, she could open the door because no key is needed [P(E|P) is close to one] but anybody else who wishes could do the same[P(E|~P) is also close to one]. At home, however, the probability of the door opening when someone other than your partner wants to come in [P(E|~P)] is close to zero, perhaps because s/he is the only one with the required key. So when you hear the door open at home, you have practical certainty of who is coming in, unless other household members also have keys. This is because, as the Bayes formula predicts, P(E|it is your partner) is high while P(E|it is not your partner) is low. Now suppose that other household members also have keys, so that when you hear the door open, in principle it could be any of them. Normally, in such cases, you still have stronger suspicions in relation to some household members than others, because you are usually aware of their plans. The plans could have changed, but usually they do not so if you are expecting one member at a certain hour while you know that other members are miles away from the house, when you hear the door open, you know who it is: the prior probability that it is the one you were expecting is very high. This probability is increased when people show up exactly when they said they would show up: while there is always a small probability that it could be someone else, this probability gets infinitely smaller that someone else shows up at exactly that time, and so on. This is why sometimes when we expect a phone call at a specific moment (e.g. a few seconds from now) we answer the phone talking directly to the person we are expecting to talk to, instead of saying 'hello' in the normal way, when we do not know who we are speaking with: the chances of receiving a call in the space of a few seconds are normally close to zero (although of course they increase as the time frame increases: the chances of receiving a call in a day are much higher, they get close to one in a week, etc.).

One statistical equivalent procedure to obtain the above predictions about who is opening the door is to observe the same event (the door opening) as many times as possible and eventually calculate the number of times it was your partner or someone else, using that probability to predict who it will be next time. What is missing in this procedure is that contextual information is not taken into account. For example, even if 95 per cent of the times the house door opens it is your partner, next time it might not necessarily be

her/him if you have just learned that s/he lost the keys. The prior probability of the event [P(E|it is your partner)], which is missing from traditional statistical methods based on frequentist probability, becomes suddenly very low, and with it the posterior probability, too [P(it is your partner|E)].

13 One type of straw-in-the-wind test is the judgement about an intervention provided by a stakeholder who has stakes/interests in providing a positive judgement. If s/he indeed provides positive judgement, this is what we expected and does not really change our prior confidence in the theory.

14 Returning to the opening door example: what evidence do we look for in order to guess who opened the door? First of all, if the door requires a key to open, anyone who does not have access to that key fails the hoop test. The solid theory in this case is that in order to open the door you must have the key: it is a necessary condition. The smoking gun in this case might be hearing the person coming in, taking his shoes off, and dropping the shoes in what appears to be the same place that your partner drops theirs in: you might not actually see it, but hear the same sounds that you always hear when this happens. What is the chance that these sounds have been produced by someone other than your partner entering the house? Very low, practically zero. You deduce that this evidence is extremely unlikely to have been produced had your partner not just come in, and you safely believe that this is exactly what has happened, long before seeing it with your eyes.

15 Factor F is a specific factor or mechanism that seems plausible to the evaluator and has passed the hoop test.

16 There is an earlier step, which we might call P0, where we build our prior confidence on the causal mechanism on the basis of existing knowledge and literature, which is not a proper test in the sense defined here and therefore is not addressed in detail.

17 These might be both supporting factors (i.e. factors that are necessary for the intervention to work) and also more independent, external factors.

18 The CA steps up to and including step four are not completely unrelated to PT. They are useful to estimate the prior probability of the ToC holding, which we do not address in detail here.

19 More specific questions can be added, making confirmation of the mechanism stronger.

20 The more alternatives are tested, the higher the chances that the remaining theory built through combining all the surviving factors will be strong enough to pass the doubly-decisive test.

References

Beach, D. and Pedersen, R.B. (2012) *Process-Tracing Methods: Foundations and Guidelines*, Ann Arbor MI: University of Michigan Press

Beach, D. and Pedersen, R.B. (2011) 'What is Process Tracing Actually Tracing? The Three Variants of Process Tracing Methods and their Uses and Limitations', paper prepared for presentation at the American Political Science Association annual meeting, Seattle, Washington, 1–4 September 2011

Befani, B. (2013) 'Between Complexity and Generalization: Addressing Evaluation Challenges with QCA', *Evaluation* 19.3: 269–83

Befani, B. (2012) 'Models of Causality and Causal Inference', in E. Stern, N. Stame, J. Mayne, K. Forss, R. Davies and B. Befani (eds), *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, London: Department for International Development

Bennett, A. (2010) 'Process Tracing and Causal Inference', in H. Brady and D. Collier (eds), *Rethinking Social Inquiry*, Lanham MD: Rowman and Littlefield

Bennett, A. (2008) 'Process Tracing: A Bayesian Perspective', in J.M. Box-Steffensmeier, H.E. Brady and D. Collier (eds), *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press: 702–21

Brady, H.E. (2002) 'Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory', paper presented at the Annual Meetings of the Political Methodology Group, University of Washington, Seattle, Washington, 16 July 2002

Collier, D. (2011) 'Understanding Process Tracing', *Political Science and Politics* 44.4: 823–30

Collier, D.; Brady, H.E. and Seawright, J. (2010) 'Sources of Leverage in Causal Inference: Towards an Alternative View of Methodology, in D. Collier and H.E. Brady (eds), *Rethinking Social Inquiry*, Lanham MD: Rowman and Littlefield

Duflo, E.; Glennerster, R. and Kremer, M. (2008) 'Using Randomization in Development

Economics Research: A Toolkit', in T. Schultz and J. Strauss (eds), *Handbook of Development Economics* Vol 4, Amsterdam and New York NY: North Holland

George, A.L. and Bennett, A. (2005) *Case Studies and Theory Development in the Social Sciences*, Cambridge MA: MIT Press

Gertler, P.J.; Martinez, S.; Premand, P.; Rawlings, L.B. and Vermeersch, C.M.J. (2011) *Impact Evaluation in Practice*, Washington DC: World Bank

Glennan, S.S. (1996) 'Mechanisms and the Nature of Causation', *Erkenntnis* 44: 49–71

HM Treasury (2011) *The Magenta Book: Guidance for Evaluation*, London: HM Treasury

Leeuw, F. and Vaessen, J. (2009) *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*, Washington DC: Network of Networks on Impact Evaluation (NONIE)

Mackie, J.L. (1974) *The Cement of the Universe: A Study of Causation*, Oxford: Oxford University Press

Mayne, J. (2012a) *Making Causal Claims*, Brief 26, Institutional Learning and Change Initiative

Mayne, J. (2012b) 'Special Issue: Contribution Analysis', *Evaluation* 18.3

Mayne (2011) 'Contribution Analysis: Addressing Cause and Effect', in R. Schwartz, K. Forss and M. Marra (eds), *Evaluating the Complex*, New Brunswick NJ: Transaction Publishers: 53–96

Mayne, J. (2008) *Contribution Analysis: An Approach to Exploring Cause and Effect*, Brief 16, Institutional Learning and Change Initiative

Mayne, J. (2001) 'Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly', *Canadian Journal of Program Evaluation* 16.1: 1–24

Mayne, J. and Stern, E. (2013) *Impact Evaluation of Natural Resource Management Research Programs: A Broader View*, ACIAR Impact Assessment Series Report 84, Canberra: Australian Centre for International Agricultural Research

Mill, John Stuart (1843) *A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*, 2 vols, London

OECD (2005) *Teachers Matter: Attracting, Development and Retaining Effective Teachers*, Paris: OECD

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*, London: Sage Publications

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*, London: Sage Publications

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage Publications

Ragin, C.C. (1989) *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*, London: University of California Press

Rihoux, B. and Ragin, C.C. (2009) *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*, Thousand Oaks CA: Sage Publications

Rohlfing, I. (2013) 'Comparative Hypothesis Testing Via Process Tracing', *Sociological Methods & Research*, 15 October

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R. and Befani, B. (eds) (2012), *Broadening the Range of Designs and Methods for Impact Evaluations*, DFID Working Paper 38, London: Department for International Development

Van Evera, S. (1997) *Guide to Methods for Students of Political Science*, New York NY: Cornell University Press

Weiss, C.H. (1997a) *Evaluation: Methods for Studying Programs and Policies*, 2nd ed., London: Prentice Hall

Weiss, C.H. (1997b) 'Theory-based Evaluation: Past, Present, and Future', *New Directions for Evaluation* 1997.76: 41–55

Weiss, C.H. (1995) 'Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families', in J. Connell, A.C. Kubisch, L.B. Schorr and C.H. Weiss (eds), *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*, New York NY: The Aspen Institute